National Aeronautics and Space Administration

Hadoop for High-Performance Climate Analytics

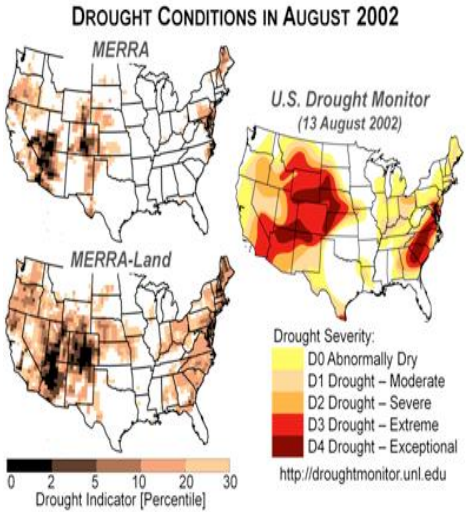
Use Cases and Lessons Learned

Glenn Tamkin (NASA/CSC)

Team: John Schnase (NASA/PI), Dan Duffy (NASA/CO), Hoot Thompson (PTP), Denis Nadeau (CSC), Scott Sinno (PTP)

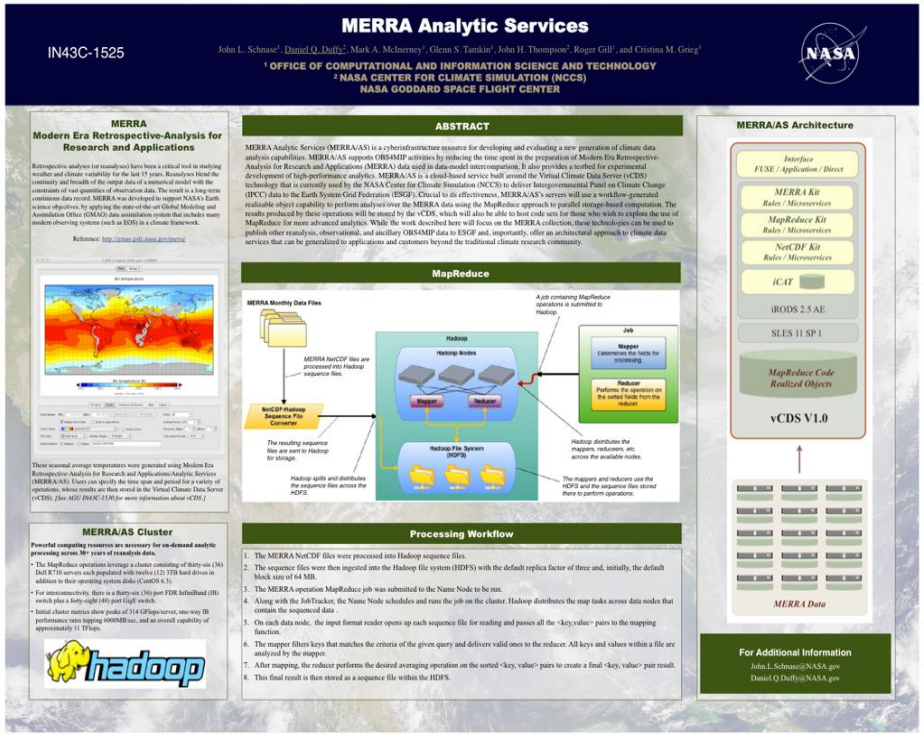
Overview

- Scientific data services are a critical aspect of the NASA Center for Climate Simulation's mission (NCCS). Modern Era Retrospective-Analysis for Research and Applications Analytic Services (MERRA/AS) …
  - Is a cyber-infrastructure resource for developing and evaluating a next generation of climate data analysis capabilities
  - A service that reduces the time spent in the preparation of MERRA data used in data-model inter-comparison



Vision

- Provide a test-bed for experimental development of high-performance analytics
- Offer an architectural approach to climate data services that can be generalized to applications and customers beyond the traditional climate research community



MERRA A/S Background

- Initially evaluated MapReduce and the Hadoop Distributed File System (HDFS) on representative collections of observational and climate data (MERRA)
  - Focused on a small set of canonical operations such as, average, minimum, maximum, and standard deviation operations over a given temporal and spatial extent
  - Built a cluster with available hardware (then acquired a custom cluster)
  - Implemented a prototype to process the data via MapReduce
  - Captured metrics and observed performance improvements as the number of data nodes and block sizes increase

Project Details

- MERRA/AS…
  - Leverages the Hadoop/MapReduce approach to parallel storage-based computation.
  - Uses a workflow-generated approach to perform analyses over the MERRA data
  - Introduces a generalized application programming interface (API) and web service that exposes reusable climate data services.

Why HDFS and MapReduce ?

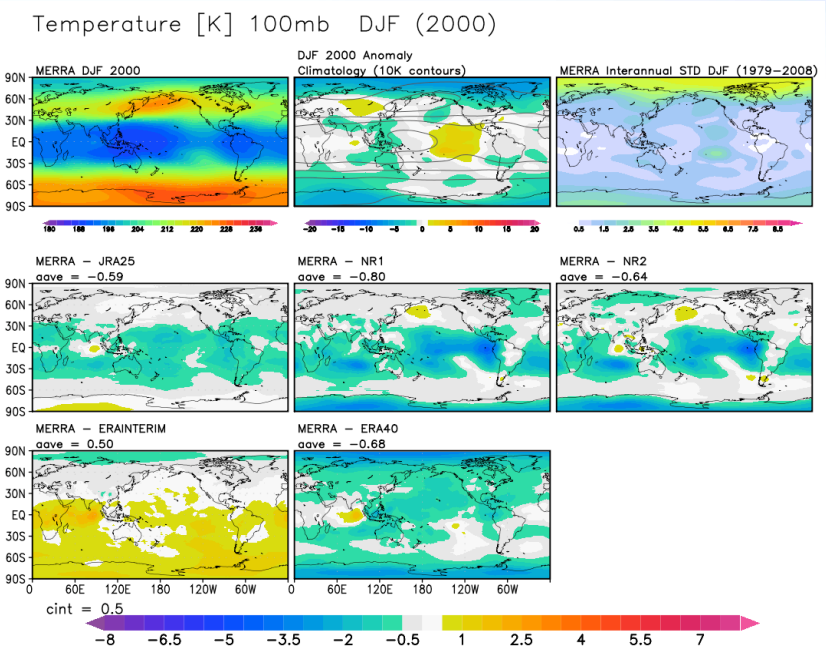
- Software framework to store large amounts of data in parallel across a cluster of nodes
  - Provides fault tolerance, load balancing, and parallelization by replicating data across nodes
  - Co-locates the stored data with computational capability to act on the data (storage nodes and compute nodes are the same – typically)
  - A MapReduce job takes the requested operation and maps it to the appropriate nodes for computation using specified keys

Who uses this technology?
- Google
- Yahoo
- Facebook

Many PBs and probably even EBs of data.

Initial Use Case

- Create a time-based average over the monthly means for specific variables
- This example shows a seasonal average of temperature for the winter of 2000
- Focused on reducing the time spent in the preparation of reanalysis data used in data-model inter-comparison, a long sought goal of the climate community



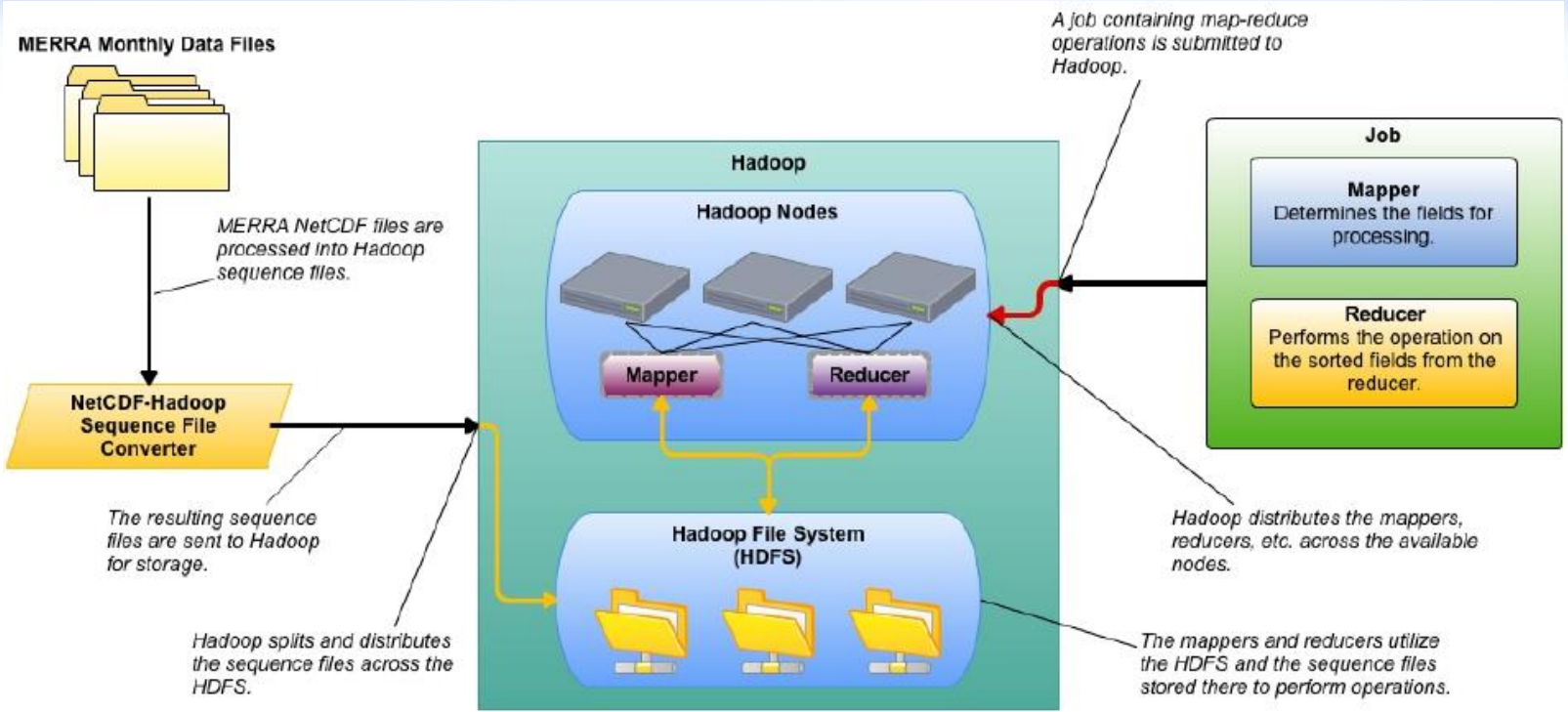
MERRA Data

- The GEOS-5 MERRA products are divided into 25 collections: 18 standard products, 7 chemistry products
- Comprise monthly means files and daily files at six-hour intervals running from 1979 – 2012
- Total size of netCDF MERRA collection in a standard filesystem is ~80 TB
- One file per month/day produced with file sizes ranging from ~20 MB to ~1.5 GB

Name	Description	Size Gbytes/day // Tbytes
const_2d_asm_Nx	Constant fields	
inst6_3d_ana_Nv	Analyzed fields on model layers	0.452
inst6_3d_ana_Np	Analyzed fields at pressure levels	0.291
inst3_3d_asm_Cp	Basic assimilated fields from IAU corrector	0.231
tavg3_3d_cld_Cp	Upper-air cloud related diagnostics	0.075
tavg3_3d_mst_Cp	Upper-air diagnostics from moist processes	0.056
tavg3_3d_trb_Cp	Upper-air diagnostics from turbulence	0.147
tavg3_3d_rad_Cp	Upper-air diagnostics from radiation	0.088
tavg3_3d_tdt_Cp	Upper-air temperature tendencies by process	0.191
tavg3_3d_udt_Cp	Upper-air wind tendencies by process	0.224
tavg3_3d_qdt_Cp	Upper-air humidity tendencies by process	0.166
tavg3_3d_odt_Cp	Upper-air ozone tendencies by process	0.083
tavg1_2d_slv_Nx	Single-level atmospheric state variables	0.285
tavg1_2d_flx_Nx	Surface turbulent fluxes and related quantities	0.267
tavg1_2d_rad_Nx	Surface and TOA radiative fluxes	0.189
tavg1_2d_lnd_Nx	Land related surface quantities	0. 146
tavg1_2d_int_Nx	Vertical integrals of tendencies	1.500
inst1_2d_int_Nx	Vertical integrals of quantities	0.115
TOTAL		4.506 // 49.6

Name	Description	Size (Gbytes)
const_2d_chm_Fx	2-D invariants on chemistry grid	
tavg3_3d_chm_Fv	Chemistry related 3-D at model layer centers	0.329
tavg3_3d_chm_Fe	Chemistry related 3-D at model layer edges	0.166
tavg3_2d_chm_Fx	Chemistry related 2-DSingle-level	0.020
tavg3_3d_chm_Nv	Accumulated transport fields at layers	0.915
tavg3_3d_chm_Ne	Accumulated transport fields at edges	0.469
inst3_3d_chm_Ne	Instantaneous fields for off-line transport	0.050
TOTAL CHEM		1.949 // 21.44

Map Reduce Workflow

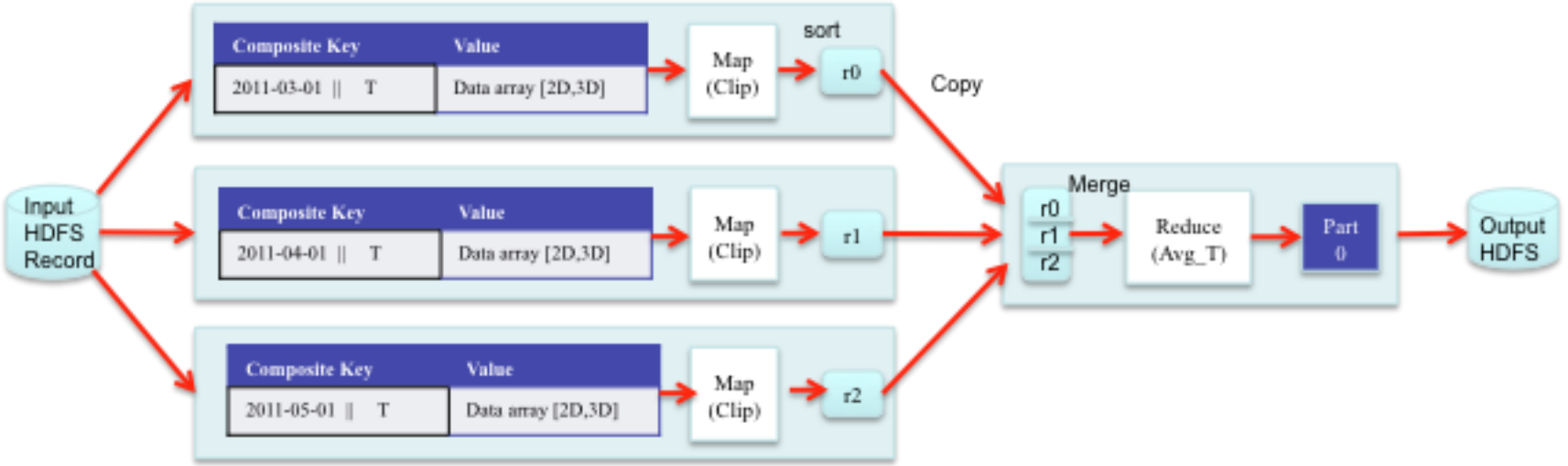


Ingesting MERRA data into HDFS

- Option 1: Put the MERRA data into Hadoop with no changes
  - » Would require us to write a custom mapper to parse
- Option 2: Write a custom NetCDF to Hadoop sequencer and keep the files together
  - » Basically puts indexes into the files so Hadoop can parse by key
  - » Maintains the NetCDF metadata for each file
- Option 3: Write a custom NetCDF to Hadoop sequencer and split the files apart
  - » Breaks the connection of the NetCDF metadata to the data
- Chose Option 2

Sequence File Format

- During sequencing, the data is partitioned by time, so that each record in the sequence file contains the timestamp and name of the parameter (e.g. temperature) as the composite key and the value of the parameter (which could have 1 to 3 spatial dimensions)



Data Set Descriptions

- **Two data sets**
  - MAIMNPANA.5.2.0 (instM_3d_ana_Np) – monthly means
  - MAIMCPASM.5.2.0 (instM_3d_asm_Cp) – monthly means
- **Common characteristics**
  - Spans years 1979 through 2012…..
  - Two files per year (hdf, xml), 396 total files
- **Sizing**

Type	Raw Total (GB)	Sequenced Total (GB)	Raw File (MB)	Sequenced File (MB)	Sequence Time (sec)
MAIMNPANA	84	224	237	565	30
MAIMCPASM	48	119	130	300	15

Seasonal Averages – Operational Cluster

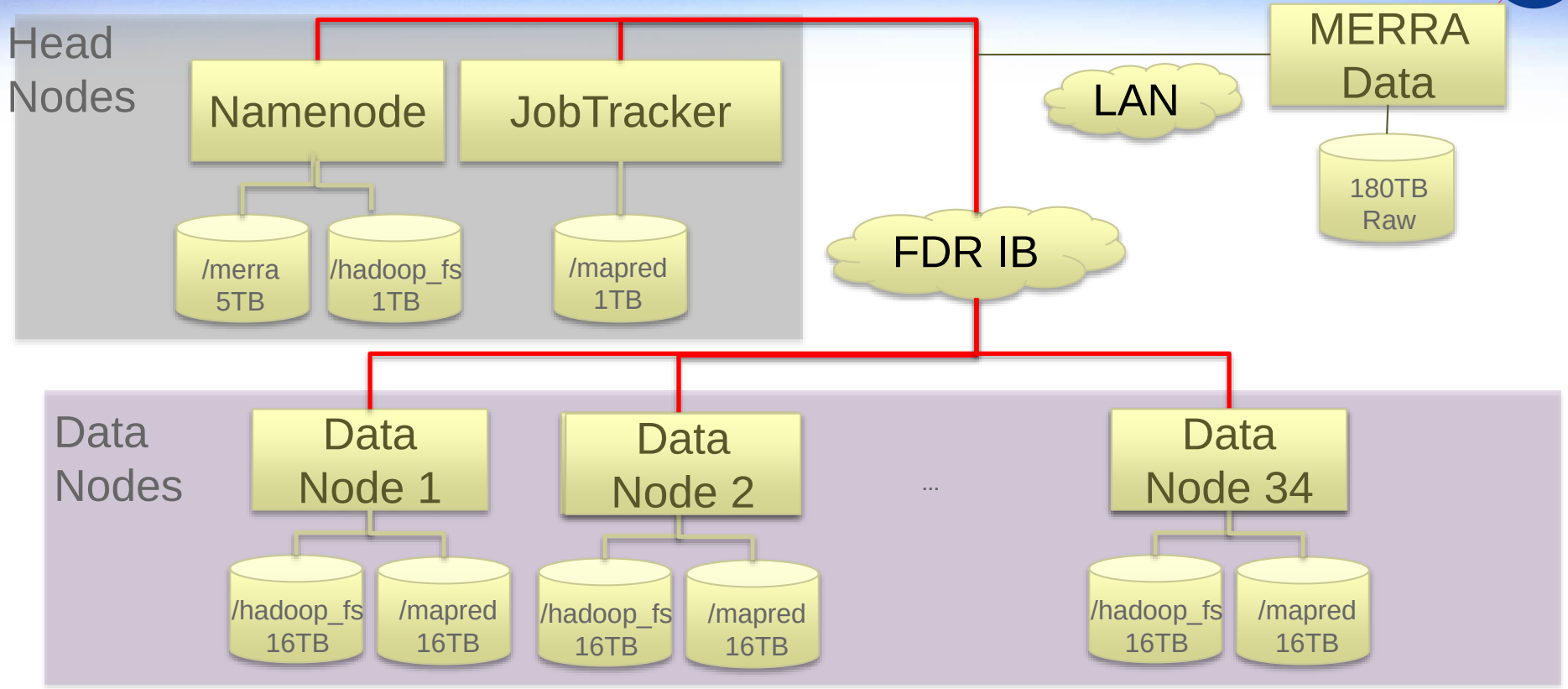
- **MAIMNPANA.5.2.0 (sec)**

Years	Period	HDFS Blocking (640MB)		
		Test	Operational	Speedup
1	2001	89.1	32.4	2.8
10	2001 - 2010	475.4	128.8	3.7
20	1991 - 2010	1026.6	245.2	4.2
All	1979 - 2011	1520.0	404.7	3.8

- **MAIMCPASM.5.2.0 (sec)**

Years	Period	HDFS Blocking (640MB)		
		Test	Operational	Speedup
1	2001	65.4	18.5	3.5
10	2001 - 2010	205.0	38.7	5.3
20	1991 - 2010	358.1	79.8	4.5
All	1979 - 2011	545.6	110.8	4.9

MERRA Cluster Components



Operational Node Configurations

Configuration	Bare1
Node	Dell R720
Processor Type	Intel Sandy Bridge
Processor Number	E5-2670
Processor Speed	2.60 GHz
Cores per Socket	8
Number of Sockets	2
Cores per Node	16
Main Memory	32 GB
Storage	12 by 3 TB drives = 36 TB RAW
Interconnect	Mellanox MT27500 FDR IB
Operating System	Centos 6.3
Kernel	2.6.32-279.5.1
Hadoop	0.20.2
java-6-sun	1.6.0_24

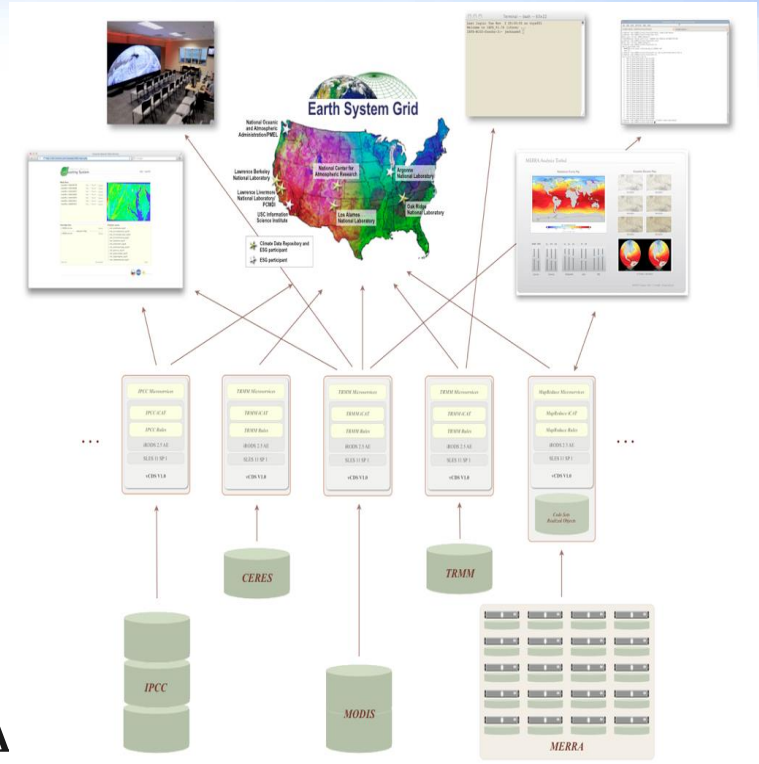
Open Source Tools

- Using Cloudera (CDH), the open source enterprise-ready distribution of Apache Hadoop.
- Cloudera is integrated with configuration and administration tools and related open source packages, such as Hue, Oozie, Zookeeper, and Impala.
- Cloudera Manager Free Edition is particularly useful for cluster management, providing centralized administration of CDH.

Customer Connections

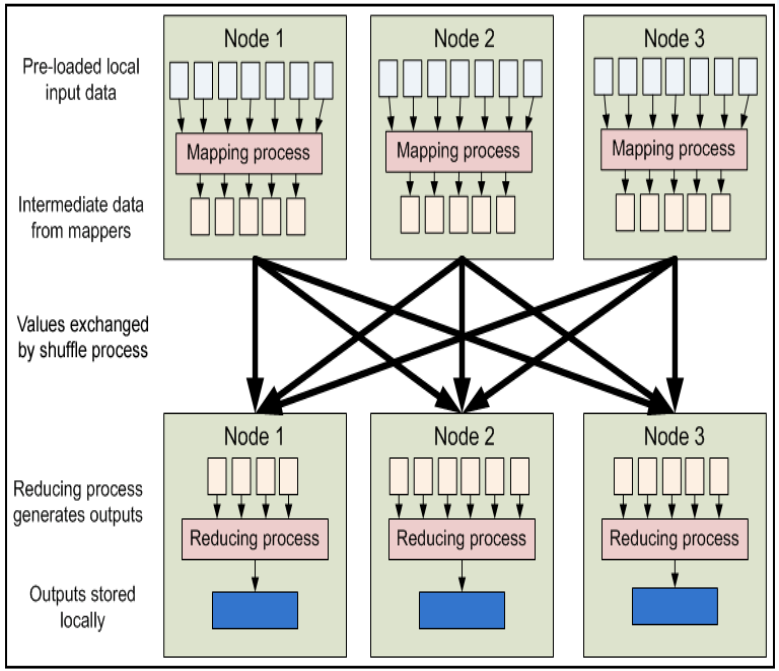
- NASA ASP A.35 Wildland Fires RECOVER project.
- NSF DataNet Federation Consortium
- SIGClimate

- Others include: GSFC / LARC iRODS Testbed, CSC Climate Edge product line, Applied Science and Terrestrial Ecology Program climate adaptation projects, Direct Readout Laboratory Climate Data Records (CDRs), and NCA modelers



Next Steps

- Tune the MapReduce Framework
- Identify potential performance optimizations (e.g., modify block size, tweak I/O config
- Complete canonical operations (e.g., add mappers/reducers)
- Try different ways to sequence the files
- Experiment with data accelerators



Conclusions and Lessons Learned

- Design of sequence format is critical for big binary data
- Configuration is key…change only one parameter at a time
- Big data is hard, and it takes a long time….
- Expect things to fail – a lot
- Hadoop craves bandwidth
- HDFS installs easy but optimizing is not so easy
- Not as fast as we thought … is there something in Hadoop that we don't understand yet
- It's all still cutting edge to a certain extent
- Ask the mailing list or your support provider